BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 9(1)

MethFinder - A software package for prediction of human tissue-specific methylation status of CpG islands

Priyanka James[$], Reshmi Girijadevi[$], Sona Charles & M Radhakrishna Pillai*

Cancer Research Program, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, Kerala-695014, India; M Radhakrishna Pillai – Email: mrpillai@rgcb.res.in; *Corresponding author
$ - Authors contributed equally

Received December 06, 2012; Accepted December 10, 2012; Published January 09, 2013

Abstract:
DNA methylation, the highly studied epigenetic mechanism which is involved in the regulatory events of various cellular processes like chromatin structure modifications, chromosomal inactivation, gene expressional patterns, embriyonic developments and transcriptional modification etc. Various high throughput techniques evolved for direct detection of methylation actions as well as information across the entire region. However, despite high throughput technological advances in experimental field, the development of software tools that has been dedicated to the prediction of epigenetic information from specific genome sequences is warranted. To this end we developed a tissue specific classifier **MethFinder** based on the frequency of novel sequence patterns across nine human tissues that was capable of discriminating methylation prone and methylation resistant CpG islands with an overall accuracy of 93%.

Availability: MethFinder is freely available at www.rgcb.res.in / methfinder

Key Words: DNA methylation, CpG islands, Support Vector Machines (SVM).

Background:
High levels of epigenetic systems such as DNA methylation, histone modification and chromatin remodelling tightly regulate gene specificity in mammals **[1].** DNA methylation, is the widely studied epigenetic modification and has a critical role in tissue-specific gene expression in mammals. Computational approaches for detection of methylation events would be a complimentary aid for expensive and laborious experimental analysis. Genome-wide DNA methylation studies show that methylation status is tissue specific and possess sequence correlations **[2, 3].** Recently some studies revealed evolutionary conservation of tissue-specific methylation in human tissues by using BAC microarrays **[3].** Both experimental and computational comprehensive genome-wide profiles of methylated regions would significantly improve our ability to address these questions. Currently there are no tissue specific methylation tools available, thus a need for a classifier that can detect patterns across tissues and to calculate DNA methylation levels by available statistical models. To this end, we developed **MethFinder** an efficient machine learning model to unravel the pattern of DNA methylation in CpG dinucleotides using support vector machines (SVM).

Tissue-specific Sequence data sets
The tissue-specific non-redundant cytosine methylation data were extracted from MethDB **[4]** a curated database of experimentally determined methylated DNA fragments. The database contains a total of 5382 methylation patterns from

various sources ranging from plants to humans **[5].** In-house Python script was used to download tissue specific methylation patterns of Homo sapiens from MethDB. We incorporated CpG islands predicted by the CpG cluster algorithm. For studying the effect of flanking sequence features, we split the sequences into overlapping fragments of fixed window size. Fragments with a methylated cytosine in the center were considered as Methylation prone, where as fragments with non-methylated cytosine in the center were considered as Methylation resistant.

Pattern Detection and classification

To detect overrepresented sequence motifs in the flanking regions, we used the Multiple Em for Motif Elicitation (MEME Suite version 4.3.0) **[6].** Twenty best-fit motifs were obtained for each sequence set (Methylation prone and Methylation resistant) for individual tissues, for all window size using the ZOOPs model (zero or one occurrence per sequence) with default parameters. When submitted to MEME, datasets with increasing window size from 59 to 79 show the presence of motifs for nine tissues (Blood, Brain, Kidney, Liver, Lung, Muscle, Pancreas, Prostate and Skin). For each sequence, MAST a motif alignment program **[7]** determines the best match in the sequence to each motif. The frequency and position of all motif hits with a goodness-of-fit (P < 0.000001) were extracted using custom Perl scripts. The percentages of occurrence of each motif between the methylation prone and methylation resistant data sets were calculated for the datasets with window size from 59 to 79 for all tissues. A Student's t-test was used to compare the frequency of occurrence of each motif between two datasets and P-value below P>0.001 as considered as not significant between the methylation prone and methylation resistant data sets.

Support vector machine (SVM) parameter optimization and calculation

For optimization, we developed training data sets of n samples, (x_1,x_2,.....xi,.....x_n), where x_i are vectors of d features and known labels for each vector {y1,y2,y3, .. .,yi, .. .,yn}, indicating whether the fragment is methylation prone or methylation resistant (yi{+1, -1}**) (see supplementary material for equation and explanation).** We used the software LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) for the implementation of the SVM algorithm, and adopted the linear kernel function by a grid search script available in the LibSVM package, with 10 fold cross validation.

Sensitivity (SE), specificity (SP), accuracy (ACC) and Matthew correlation coefficient (MCC) of the SVMs to assess classification performance were estimated using the following equations. We calculated the expressions for SP, SE, ACC and MCC using Eqs. (2– 5) **(see supplementary material).**

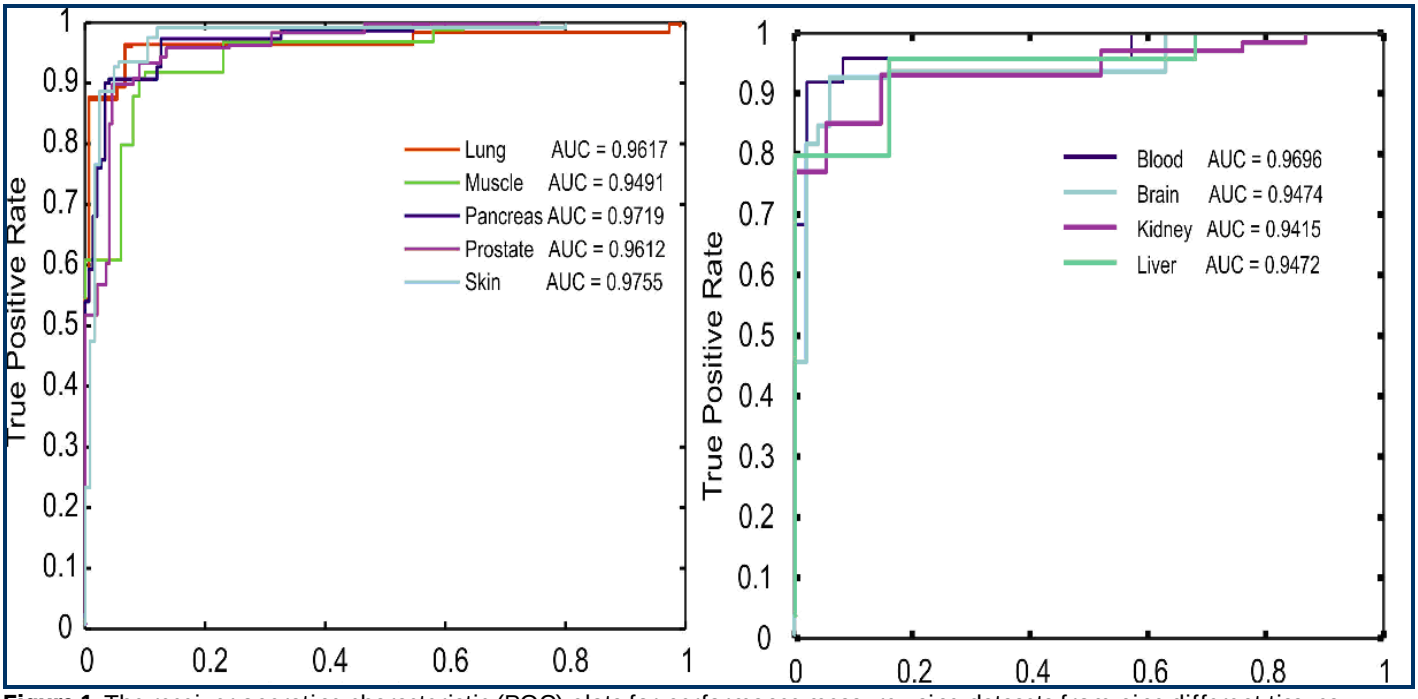

Figure 1: The receiver operating characteristic (ROC) plots for performance measure using datasets from nine different tissues

Software Performance

We trained the SVM classifier with training sets from nine different tissues (supplementary table 1) and tested its performance on a corresponding test set from individual tissues. The training set was randomly selected from individual tissues with specific window length (59, 69, and 79 bp). For each window length, this experiment was repeated with random selections of training and test sets. The best classification accuracy was observed for a window size of 69, where the best balance between specificity (0.97) and sensitivity (0.89) were also observed with the highest value for MCC (0.86) **Table 1 (see supplementary material).** Performance of the classifiers was also evaluated by forming receiver operating characteristic (ROC) curves **(Figure 1).** Here we used motif-based sequence analysis tools coupled with classification techniques to identify DNA sequence patterns that define CpG island methylation status. This study serves as proof-of-principle that the epigenetic state of a genomic region can be predicted based on DNA sequences.

Acknowledgement:
This work was sponsored by Advaita Informatics, Dubai.

References:
[1] Nagae G, *Hum Mol Gen.* 2011 **20**: 14 [PMID: 21505077]
[2] Eckhardt F *et al*. *Nat Genet.* 2006 **38**: 1378 [PMID: 17072317]
[3] Illingworth RS & Bird AP, *FEBS Lett.* 2009 **583**: 1713 [PMID: 19376112]
[4] Ghosh S *et al*. *Epigenetics.* 2010 **5**: 527 [PMID: 20505344]
[5] Negre V & Grunau C, *Epigenetics.* 2006 **1**: 101 [PMID: 17965614]
[6] Hackenberg M *et al. BMC Bioinformatics.* 2006 **7**: 446 [PMID: 17038168]
[7] Bailey TL & Gribskov M, *Bioinformatics.* 1998 **14**: 48 [PMID: 9520501]

Edited by P Kangueane
Citation: James *et al*. Bioinformation 9(1): 061-064 (2013)

BIOINFORMATION open access

Supplementary material:

Support vector machine (SVM) parameter optimization and calculation

For optimization, we developed training data sets of n samples, (x_1,x_2,.....xi,.....x_n), where x_i are vectors of d features and known labels for each vector {y1,y2,y3, .. .,yi, .. .,yn}, indicating whether the fragment is methylation prone or methylation resistant (yi{+1, -1}).SVM obtains a classifier of the form

$f(x) = sign(g(x)) = sign\{\sum_{i=1}^{n} \alpha_i y_i k(xi,x) + b\}$	→	(1)

where the i and b are optimized in the training procedure with the objective of minimizing the prediction error on training data while maximizing the separation margin between the two classes. The K is a kernel function that can be regarded as a measure of the similarity between two samples.

Sensitivity (SE), specificity (SP), accuracy (ACC) and Matthew correlation coefficient (MCC) of the SVMs to assess classification performance were estimated using the following equations. We calculated the expressions for SP, SE, ACC and MCC using Eqs. (2–5), respectively,

$SE = TP/(TP + FN)$	→	(2)
$SP = \frac{TN}{TN+FP}$	→	(3)
$ACC = \frac{TP+TN}{TP+TP+FP+FN}$	→	(4)
$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}}$	→	(5)

Table 1: Average Performance measure of the Support Vector Machine (SVM) for different window sizes of nine tissues

Window size	SP	SE	ACC	MCC
59 bp	0.95	0.83	0.91	0.83
69 bp	0.97	0.89	0.93	0.86
79 bp	0.88	0.96	0.917	0.83